Face Transformer: Towards High Fidelity and Accurate Face Swapping

Kaiwen Cui, Rongliang Wu, Fangneng Zhan, Shijian Lu*
Nanyang Technological University, Singapore
{Kaiwen001, Rongliang001}@e.ntu.edu.sg, {fnzhan,Shijian.Lu}@ntu.edu.sg

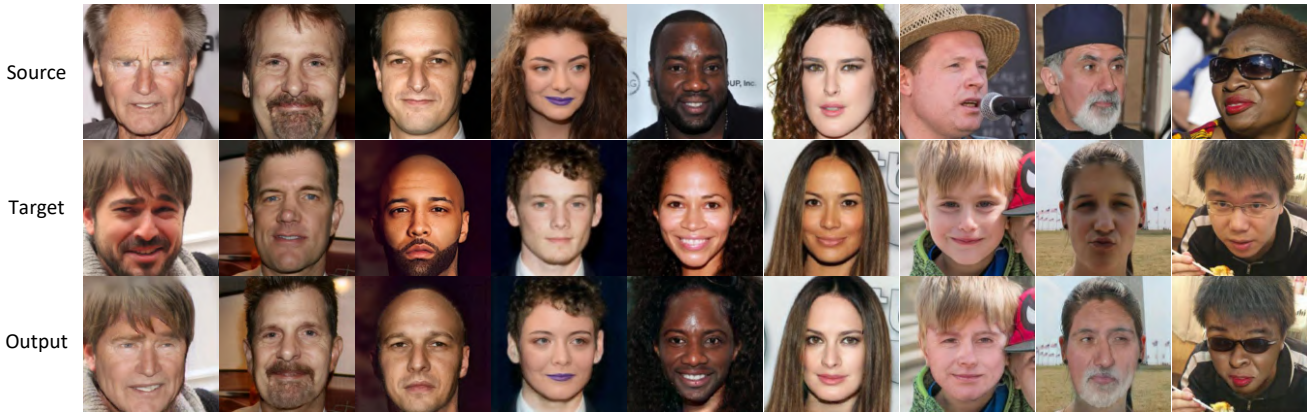

Figure 1. The proposed Face Transformer fuses the identity of $Source$ face and attributes of $Target$ faces according to the semantic-aware correspondence that is constructed by using a transformer architecture. It produces high-fidelity and accurate face swapping over multiple public datasets including CelebA-HQ [21], VGGFace [29] and FFHQ [13] (Sample images in columns 1-3, 4-6, and 7-9 are from CelebA-HQ, VGGFace, and FFHQ, respectively).

Abstract

Face swapping aims to generate swapped images that fuse the identity of source faces and the attributes of target faces. Most existing works address this challenging task through 3D modelling or generation using generative adversarial networks (GANs), but 3D modelling suffers from limited reconstruction accuracy and GANs often struggle in preserving subtle yet important identity details of source faces (e.g., skin colors, face features) and structural attributes of target faces (e.g., face shapes, facial expressions). This paper presents Face Transformer, a novel face swapping network that can accurately preserve source identities and target attributes simultaneously in the swapped face images. We introduce a transformer network for the face swapping task, which learns high-quality semantic-aware correspondence between source and target faces and maps identity features of source faces to the corresponding region in target faces. The high-quality semantic-aware correspondence enables smooth and accurate transfer of source identity information with minimal modification of target shapes and expressions. In addition, our Face Transformer incorporates a multi-scale transformation mechanism for preserving the rich fine facial details. Extensive experiments show that our Face Transformer achieves superior face swapping performance qualitatively and quantitatively.

1. Introduction

Face Swapping aims to generate new face images that combine the source faces' identities which include skin colors, face features, makeups and etc., as well as the target faces' attributes that include head poses, head shapes, facial expressions, backgrounds, etc. Automated and realistic face swapping has attracted increasing interest in recent years thanks to its wide range of applications in different areas, such as movie composition, computer games and privacy protection, etc. However, it remains a challenging task to accurately and realistically extract and fuse identity information from source faces and attribute features from target faces.

Most existing face swapping methods can be broadly classified into two categories which exploit 3D face models

and generative adversarial networks (GANs) [10], respectively. Earlier works [19, 27] employ 3D models to deal with the pose and perspective differences between source and target faces, which estimate 3D shapes of source and target faces and use the estimated 3D shapes as proxy for face swapping. However, 3D based methods suffer from limited accuracy in 3D reconstruction which tend to generate various distortions and artefacts in the swapped face images. Inspired by the great success of GANs [10,23,31,48], several works [1, 17, 18, 24–26] employ generative models and have achieved very impressive face swapping performance. Though GAN-based networks can generate high-fidelity swapping, they still struggle in preserving subtle yet important identity details of source faces (in skin colors, face features, makeups, etc.) and structural attribute features of target faces (in face shapes, facial expressions, etc.).

This paper presents Face-Transformer, an innovative face swapping network that achieves superior face swapping by accurate preservation of identity details of source faces and structural attributes of target faces. The superior swapping performance is largely attributed to a transformer architecture we introduce into the face swapping task. Specifically, we employ the transformer to build up semantic-aware correspondence between source and target faces with which the identity features of source faces can be mapped to the corresponding region of target faces smoothly and accurately. The construction of semantic-aware correspondence can not only preserve structural attributes of target faces but also achieve high-fidelity identity of source faces especially for subtle yet important details as illustrated in Fig. 1. The proposed Face Transformer consists of three modules including a face parsing module, a face feature transformation module (FFTM) and a face generation module (FGM). Face parsing module is an off-the-shelf module* which predicts face masks to separate inner faces from the background and extracts face semantics to guide the training of transformer in FFTM. FFTM takes source face, target face and target face semantics as inputs to learn semantic-aware correspondence between two faces, and maps the identity features of source faces to the corresponding regions of target faces. FGM finally generates high-fidelity swapped face images based on the transformation of multi-scale facial features by FFTM.

The contributions of this work are threefold. *First*, we design Face-Transformer, an innovative network that achieves accurate face swapping by introducing transformer into face swapping task. The transformer learns semantic-aware correspondence between source and target faces which facilitates the feature transfer from source faces to target faces smoothly. To the best of our knowledge, this is the first work that introduces transformer architectures for the face swapping task. *Second*, we propose a novel multi-scale feature transformation strategy that helps learn more powerful feature representation and achieve more accurate face swapping. *Third*, extensive experiments show that the proposed Face Transformer achieves superior face swapping quantitatively and qualitatively.

*https://github.com/zllrunning/face-parsing.PyTorch

2. Related Works

2.1. Face Swapping

Face swapping has achieved remarkable progress in recent years. Existing face swapping methods can be broadly classified in two categories: 3D-based methods and GAN-based methods.

2.1.1 3D-Based methods

Early face swapping task involves manual intervention [3] until the introduction of automated methods [2]. However, automated methods cannot preserve face expressions well and Face2Face [34] was then proposed to transfer expressions from target faces to source faces. Face2Face works by applying 3D morphable face model (3DMM) to source faces and target faces and then transferring the expression component from target face to source face. To further preserve the occlusions, Nirkin *et al*. [27] collected data to train an occlusion-aware face segmentation network in a supervised way. However, the 3D-Based models suffer from limited accuracy in 3D reconstruction which tend to generate various distortions and artefacts in the swapped face image.

2.1.2 GAN-Based methods

Generative adversarial network [10] (GAN) has achieved great success in image generation [7, 11, 16, 40–42, 44, 45]. Leveraging the fast development of GANs, several methods [1, 17, 18, 24–26] introduce GANs for face swapping and have achieved quite impressive progress.

Deepfakes† and Faceswap ‡have achieved great success in face swapping recently, but they need to train a new model with two video sequences for each new input. To mitigate this constraint, subject-agnostic face swapping has attracted increasing interest. For example, Natsume *et al*. [25] disentangles the embedding of face and hairs and recombines them to generate swapped face. Natsume *et al*. [24] utilizes a latent space to preserve face identity in source face and appearance of hair style and background region in target face. Nirkin *et al*. [26] utilizes an occlusion-aware face segmentation network to preserve facial occlusion and Li *et*

†https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes

‡https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski

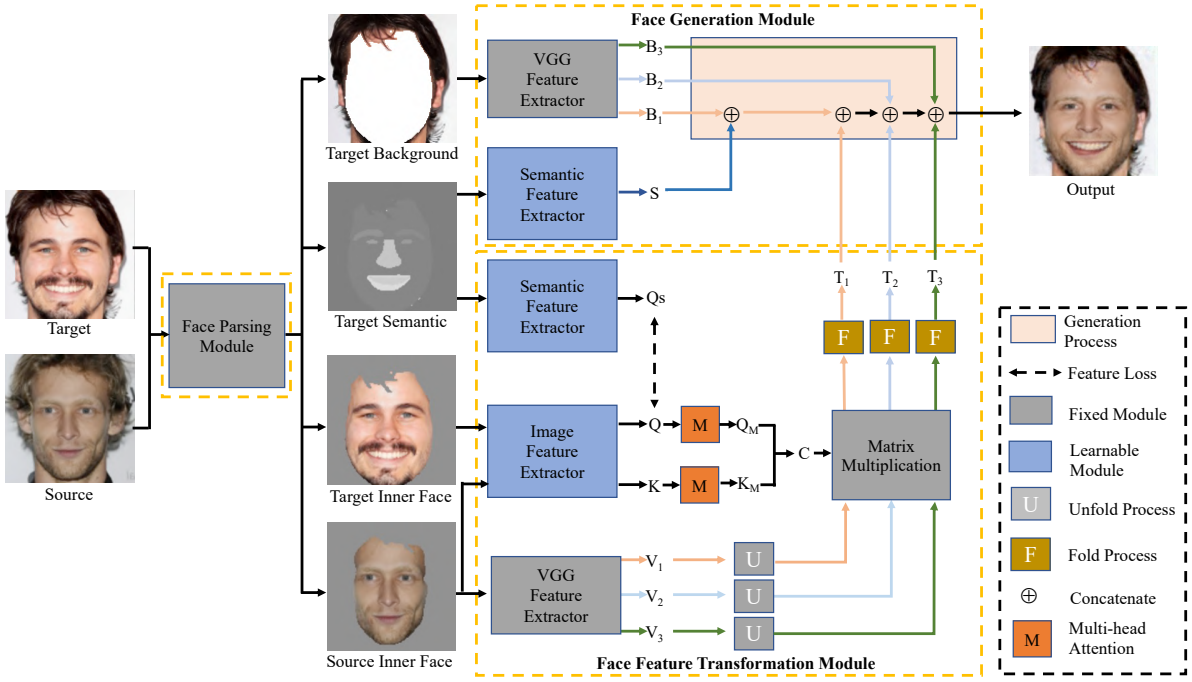

Figure 2. The architecture of the proposed Face Transformer: Given a *Source* face and a *Target* face, a *Face Parsing Module* first extracts a $Source\ Inner\ Face$, a $Target\ Inner\ Face$, a $Target\ Semantic$ map, and a $Target\ Background$, respectively. A transformer then learns a semantic-aware correspondence matrix C from source and target face features K and Q. The learned C maps multi-scale identity features V_1, V_2, and V_3 of the source face (extracted by a pre-trained $VGG\ Feature\ Extractor$) to T_1, T_2, and T_3 that are aligned with the corresponding region of the target face. Finally, T_1, T_2, and T_3 are concatenated with multi-scale background features B_1, B_2, and B_3 of the target face to generate the swapped face $Output$.

al. [18] presents a heuristic error acknowledgement refine network. Although GAN-based methods can achieve high fidelity face swapping, most of them struggle in preserving subtle yet important identity details of source faces (in skin color, face features, makeups, etc.) and structural attributes of target face (in face shape, face expression, etc.). The proposed Face Transformer learns a semantic-aware correspondence between source faces and target faces that can accurately and smoothly transfer source identity information with minimal modification of the shapes and expressions of target faces.

2.2. Transformer

Initially proposed for natural language processing tasks, transformer has recently become an emerging component that is widely applied in various vision tasks. With its superiority over convolutional neural networks (CNNs) in capturing long-distance relationship, transformer-based vision frameworks have demonstrated their effectiveness in image classification [8, 20, 35], object detection [4, 46, 49], image synthesis [6, 9, 39, 43], super-resolution [38], etc. The key in the transformer architecture is the attention mechanism [36] that models interactions between its inputs regardless of their relative position to one another.

In this work, we exploit the strong expressivity of the transformer architecture for the challenging face swapping task, where the major challenge is to construct accurate semantic-aware correspondence between source and target faces. The superior performance achieved by the proposed Face Transformer demonstrate that the transformer architecture is a natural fit for this task.

3. Methods

As illustrated in Fig. 2, the proposed Face Transformer consists of three modules including a face parsing module, a face feature transformation module (FFTM) and a face generation module (FGM). The face parsing module is an off-the-shelf face parsing model that produces face masks and face semantics, which separate inner faces from the image background and guide the learning of semantic-aware correspondence to be used in FFTM. Once the inner face and face semantics are obtained, FFTM learns the semantic-aware correspondence between source inner face and target inner face and transforms multi-scale features of source inner face to corresponding regions of target faces based on the learnt semantic-aware correspondence. The multi-scale transformed features are then progressively concatenated with features of target semantic and multi-scale features of

the target-face background for the generation swapped face images. FFTM and FGM are trained in an end-to-end manner as to be discussed in the ensuing subsections.

3.1. Face Feature Transformation Module

With the extracted source inner faces, target inner faces and target semantics, FFTM learns to map identity features of the source inner faces to target inner faces according to the semantic-aware correspondence between the source and target face features.

3.1.1 Feature extractor

FFTM has three feature extractors. The first is a VGG feature extractor which employs a pre-trained VGG-19 model to extract multi-scale features $V_1 \in \mathcal{R}^{H*W*C}$, $V_2 \in \mathcal{R}^{2H*2W*\frac{C}{2}}$, and $V_3 \in \mathcal{R}^{4H*4W*\frac{C}{4}}$ from the source inner face. The extracted features are then mapped to the corresponding regions of target face based on the learnt semantic-aware correspondence. The second extractor is a learnable image feature extractor which extracts features of source inner face (K) and target inner face (Q). The third is a learnable semantic feature extractor which extracts features from the target semantic map (Q_S) for guiding the generation of semantic-aware target inner face feature Q. The incorporation of the learnable semantic feature extractor is based on the observation that target semantics capture facial expression and shape information which is critical in face swapping. The two learnable feature extractors take similar network architecture as in [47].

3.1.2 Feature Transformation

The major component of FFTM is a transformer that first performs multi-head attention over source face features K and target face features Q. This produces the corresponding feature representation Q_M and K_M. The transformer then performs cross attention between Q_M and K_M which builds up connections between the input K and output Q.

Similar to [36], the multi-head attention over Q and K is formulated as $Q_M = [head_1, ..., head_h]W_0$ and

$$head_i = softmax(\frac{(QW_i^Q)(KW_i^K)^T}{|QW_i^Q||(KW_i^K)|}), \quad (1)$$

where W denotes learnable parameters. K_M is obtained similarly.

Different from the cross attention in typical transformers, we first employ Q_M and K_M to learn a correspondence matrix C that represents the semantic-aware correspondence between source face features and target face features. The learned semantic-aware correspondence matrix is then applied over the facial features that are extracted from the VGG feature extractor. Specifically, the semantic-aware correspondence matrix $\mathcal{C}$ can be obtained by performing dot products of each channel-wise feature in Q_M with all channel-wise feature in K_M and applying softmax over the dot-product results:

$$\mathcal{C} = softmax(\frac{Q_M K_M^T}{|Q_M||K_M|}). \quad (2)$$

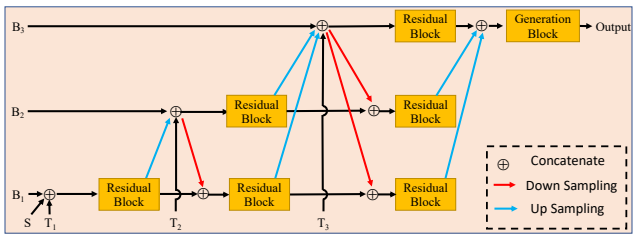

Figure 3. The face generation details: Taking semantic feature S, transformed multi-scale face features T_1, T_2, T_3 and multi-scale background features B_1, B_2, and B_3 as input, the face generator progressively and pair-wisely exchanges information among features of different scales. It handles feature scale differences by up-sampling (with an interpolation layer, a convolution layer and a ReLU layer) and down-sampling (with a convolution layer and a ReLU layer). Residual block has the same architecture as [28] and the generation block contains a convolution layer and a tanh layer.

3.1.3 Multi-scale feature transformation

We also design a multi-scale feature transformation by transforming $V_1 \in \mathcal{R}^{H*W*C}, V_2 \in \mathcal{R}^{2H*2W*\frac{C}{2}}, V_3 \in \mathcal{R}^{4H*4W*\frac{C}{4}}$ simultaneously by using the learned correspondence matrix. To resolve the spatial discrepancy across multi-scale features, we apply an unfold process U to each feature which unifies the spatial dimension of all features to be H*W. The unfolded features are then multiplied with the correspondence matrix to produce transformed features. Finally, we restore the spatial dimension of each feature by a fold process F. We design the multi-scale feature transformation with two major purposes. First, transforming multi-scale features helps preserve more fine facial details of source faces in the swapped face image. Second, learning the multi-scale feature transformation can inversely help improve the learning of the semantic-aware correspondence matrix.

3.2. Face Generation Module

Face generation module is used to fuse the transformed multi-scale features from face feature transformation module (FFTM) as well as multi-scale features of the target backgrounds (B_1, B_2, and B_3) and synthesize final high fidelity face images. To preserve the expressions and shapes of target faces, we employ another learnable semantic feature extractor to extract semantic features as input as well.

Method	ID ↓	Pose ↓	Expression ↓	Shape ↓	Quality ↑
DeepFake	38.26	4.14	43.04	66.3	0.95
FaceSwap	37.29	**2.51**	29.1	47.8	0.972
FaceShifter	38.73	2.96	27.93	51.4	0.977
Ours	**34.49**	3.13	**23.74**	**44.3**	**0.978**

Table 1. Quantitative comparisons of the proposed Face Transformer ($Ours$) with state-of-the-art methods $DeepFakes$, $FaceSwap$ and $FaceShifter$ [18] over the public dataset Faceforensics++ [32]

The detailed architecture of the face generation module is shown in Fig. 3. Inspired by [38] that exchanging information between features of different scales in generation can assist generator in learning more powerful feature representations and preserve better texture details, our face generation module progressively and pair-wisely exchanges information between features of different scales as illustrated in Fig. 3.

3.3. Training Objectives

As mentioned in Section 3.1, we employ L1 loss between Q and Q_S to preserve semantic information in Q. The feature loss $\mathcal{L}_f$ between Q and Q_S is formulated by:

$$\mathcal{L}_f = ||Q - Q_S||_1. \tag{3}$$

To synthesize visually realistic images, the finally swapped face images (f_{swap}) are generated under adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}[logD(f_{real})] + \mathbb{E}[log(1 - D(f_{swap})], \tag{4}$$

where f_{real} is real face image.

Face swapping task requires the swapped face (f_{swap}) to have the same shapes and expressions as the target face f_{tgt}. f_{swap} and f_{tgt} should therefore be semantic alike. Inspired by [12], we adopt the perceptual loss $\mathcal{L}_{perc}$ to minimize the semantic discrepancy:

$$\mathcal{L}_{perc} = ||\phi_l(f_{swap}) - \phi_l(f_{tgt})||_1. \tag{5}$$

Following [47], we choose ϕ_l to be the activation after relu4_2 layer in the VGG-19 network as this layer is highly semantic-related.

In addition, we adopt the contextual loss $\mathcal{L}_{context}$ [22] to align the face identities (in face color, texture, etc.) of the swapped face f_{swap} with the source face f_{src}:

$$\mathcal{L}_{context} = \sum_l -log(CX(\phi_l(f_{swap} * m_{tgt}), \phi_l(f_{src} * m_{src}))), \tag{6}$$

where m_{tgt} and m_{src} are the mask of target face and source face that are produced by the face parsing network. Since the background of the swapped face shall be the same as the background of the target face, we employ respective masks to the swapped faces and the source faces to filter out background information and apply the mask of the target face to the swapped face (as they are ideally the same). The details of contextual similarity CX is defined in [22]. We choose ϕ_l to be the activation after relu3_2, relu4_2 and relu5_2 as low-level features are more style-related.

The overall loss function of the proposed Face Transformer is

$$\mathcal{L} = \lambda_1\mathcal{L}_f + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{perc} + \lambda_4\mathcal{L}_{context}. \tag{7}$$

We follow settings in [47] to set λ_2 = 10, λ_3=0.001 and λ_4 = 1. We then empirically adjust λ_1 to be 5.

4. Experiments

4.1. Datasets and Settings

Similar to existing works [18, 26], we evaluate and compare the proposed Face Transformer with state-of-the-art face swapping methods DeepFakes, FaceSwap and FaceShifter [18] over the public dataset Faceforensics++ [32]. DeepFakes and FaceSwap are trained directly with Faceforensics++, while FaceShifter and the proposed Face Transformer are subject agnostic. Specifically , Faceshifter is trained by using the three datasets CelebA-HQ [21], FFHQ [13] and VGGFace [29] while the proposed Face Transformer is trained over CelebA-HQ [21] only. CelebA-HQ is a large-scale face image dataset that has 30K high-resolution face images. FFHQ consists of 70K high-quality images that contains considerable variations interms of age, ethnicity and image background. VGGFace consists of 2.6M high-resolution images that distribute over 2.6k identities.

Note Faceforensics++ [32] is a public dataset which contains 1000 videos of different identities. In our experiments, we follow [18] to evenly sample 10 frames from each video to form a test set, which consists of 10K face images.

4.2. Implementation Details

Following [18, 26], we use five-point landmarks [5] to crop and align face images. The cropped images are resized to 256 × 256 in model training. We use PyTorch [30] to implement the proposed Face Transformer. Adam optimizer [15] is adopted as the optimizer with β_1=0.5 and β_2= 0.999. The batch size is set to 4. We train the model for 20 epochs with a fixed learning rate of 0.0002.

4.3. Evaluation Metrics

We perform quantitative evaluations and comparisons with several metrics that have been used in prior research. The metrics include *ID verification*, *pose*, *expression*, *shape* and *quality*.

Figure 4. Qualitative comparisons of the proposed Face Transformer ($Ours$) with state-of-the-art methods $DeepFakes$, $FaceSwap$ and $FaceShifter$ [18] over the public dataset Faceforensics++ [32].

4.3.1 ID verification

ID verification metric examines whether the swapped face images preserve the identity information of the source faces. It is evaluated according to the Euclidean distance between the identity features of the swapped and source face images that are extracted by a pre-trained face recognition model [37]. Smaller distances indicate better identity preservation. We did not follow [26] that uses dlib to extract identity feature as [37] has much better performance in face recognition.

4.3.2 Poses

Pose metric evaluates how the swapped faces preserve the head pose of the target faces. It is computed by the Euclidean distance between head pose of target faces and the swapped faces. Following [18], the head pose of each face image is estimated by [33]. Smaller distances indicate better pose preservation.

4.3.3 Expressions

Expression metric examines how the swapped faces preserve the expressions of the target faces. It is evaluated by the Euclidean distance between 2D landmarks of the target faces and the swapped faces as in [26]. In our experiments, we use open-source software dlib [14] to detect facial landmarks. Smaller distances indicate better expression preservation.

4.3.4 Shape

Shape metric examines how the swapped faces preserve the facial shape of target faces. As state-of-the-art methods do not compare this metrics, we design it ourselves by computing the Euclidean distance between target faces' mask and swapped faces' mask. We extract face masks by using the segmentation network in [26]. Smaller distances indicate better shape similarity.

4.3.5 Quality

Quality metric evaluates the perceptual quality of the swapped face images. Following [26], we use Structural Similarity Index (SSIM) to measure the image quality. Higher SSIM indicates better image quality.

Method	ID ↓	Pose ↓	Expression ↓	Shape ↓	Quality ↑
w/o transformer	36.76	3.57	26.52	46.75	0.972
w/o multi-scale	35.16	3.87	26.63	48.81	0.974
Ours	**34.49**	**3.13**	**23.74**	**44.3**	**0.978**

Table 2. Quantitative ablation study of the proposed Face Transformer with versus w/o transformer and with versus w/o multi-scale feature transformation.

4.4. Quantitative Experiments

We quantitatively evaluate and compare the proposed Face Transformer with state-of-the-art methods DeepFakes, FaceSwap and FaceShifter [18] over dataset Faceforensics++ [32].

Table 1 shows quantitative experimental results. It can be seen that the proposed Face Transformer outperforms the state-of-the-art in terms of *ID verification*, *shape*, *expression* and *quality*, demonstrating its clear superiority in high-fidelity and accurate face swapping. However, the proposed Face Transformer does not perform the best in pose metric (3.13 v.s. 2.51) that measures the difference between the predicted 3D pose vector of the target faces and that of the swapped faces. The lower pose accuracy is largely due to the fact that we follow [18] to estimate 3D pose vectors based on image intensities [33]. However, the proposed Face Transformer preserves better skin colours of source faces in the swapped face which actually enlarges the intensity difference between the target faces and the swapped faces and further degrades the performance under the current pose metric.

4.5. Qualitative Experiments

Fig. 4 shows qualitative experimental results. As DeepFakes and FaceSwap first synthesize inner face regions and blend them with the target face backgrounds for swapped faces, they tend to generate blending inconsistency in the synthesized faces as illustrated in the sample images in column 1, 2, 3, 6, and 7. FaceShifter [18] extracts identity features from source face and attribute features from target face, and adopts attention mechanism to adaptively integrate them for face swapping. However, the strong dependency on the attention mechanism often misleads the swapping generation where the identity features of swapped faces become deviated from that of source faces such as skin colours (in all sample images) and nose shapes (in sample images in column 1, 2, 4, 5). Different from FaceShifter [18], the proposed Face Transformer first learns the semantic-aware correspondence between source face features and target face features and explicitly maps identity features of source faces to the corresponding regions of target faces to get the transformed features. The transformed features and the background features are then fused to directly generate the finally swapped faces. As the face features have already been mapped to the correct regions, the generation process becomes much easier and can achieve more accurate face swapping. It also removes the blending operation and hence eliminates the blending inconsistency issue effectively.

The proposed Face Transformer thus better preserve the identity of the source faces (e.g., the skin colors in sample images in column 1-5, the face features such as eyes, noses and mouth shapes in sample images in column 1, 2, 5, and makeups in sample images in column 1, 4, 5) and the attributes of the target faces (e.g., the mouth-opening degree in sample image in column 1, 4). We also evaluated the proposed Face Transformer over another two datasets FFHQ [13] and VGGFace [29] by directly apply the trained model to the two datasets without fine-tuning. As illustrated in Fig 1, the proposed Face Transformer can handle face images under various conditions such as large head poses and different illumination conditions. This further demonstrates the effectiveness of our Face Transformer.

4.6. Ablation Study

We perform two ablation studies quantitatively and qualitatively to demonstrate the effectiveness of our designed transformer and multi-scale feature transformation.

4.6.1 W/o transformer

We first compare the proposed Face Transformer with a baseline model that does not employ transformer. Similar to our Face Transformer, it uses the face parsing module to separate the background and inner face, extract multi-scale features from source inner face and target background, and feeds the extracted features to our face generation module. The first row in table 2 shows quantitative results. It can be seen that the proposed Face Transformer performs clearly better in all metrics. The better results is largely because our designed transformer accurately maps the features of source faces to their corresponding regions of target faces which improves the generation of swapped faces.

4.6.2 W/o multi-scale

We also compare the proposed Face Transformer with a model without using our proposed multi-scale transformation. The model uses the same transformer design but it only transforms feature V_1 in face generation. As shown in Table 2, the proposed Face Transformer performs better consistently. Without the proposed multi-scale transformation, the fidelity of swapped faces is degraded clearly. We conclude the reason of better performance with multi-scale in two aspects. First, multi-scale can preserve more fine facial details. Second, learning the multi-scale feature transformation can inversely help improve the learning of the semantic-aware correspondence matrix.

Figure 5. Qualitative illustration of the proposed Face Transformer with the same source face but different target face. Every three columns form a group of images with the same source face but different target faces. It can be observed that the proposed Face Transformer produces high-fidelity and accurate face swapping.

Figure 6. Qualitative illustration of the proposed Face Transformer with the same target face but different source faces. Every three columns form a group of images with the same target face but different source faces.

4.7. Discussion

4.7.1 Swapping with the same source faces

Fig. 5 shows face swapping by the proposed Face Transformer where the same source face is swapped to multiple target faces. It can be seen that the Face Transformer successfully transfers identity features of source faces to the corresponding regions of target faces while preserving the correct poses and expressions of the target faces (while target faces have very different expressions and poses).

4.7.2 Swapping with the same target face

Fig. 6 shows that the proposed Face Transformer can swap different source faces to the same target faces realistically. Although the skin colours of source faces are very different, our swapped faces can preserve the skin color while maintaining its compatibility with the background by adjusting the background (e.g., the color of the neck are adjusted in the first sample). This unique feature is largely attributed to the transformer design in our Face Transformer which maps the identity features of source inner face to the corresponding regions of target faces before the generation of swapped faces. The feature transformation facilitates the learning of generation model and enables it to generate high-fidelity and realistic face swapping.

5. Conclusion

In this paper, we present a novel face swapping framework named Face Transformer for accurately preserving source identities and target attributes simultaneously in the swapped face images. We introduce transformer network for the face swapping task, which learns high-quality semantic-aware correspondence between source and target faces and accurately integrates source identity information with target attributes. In addition, our Face Transformer exploits a multi-scale transformation mechanism for preserving more fine facial details. Extensive experiments show that the proposed Face Transformer can generate accurate and realistic face swapping results.

Acknowledgement

This study is funded by the Ministry of Education Singapore, under the Tier-1 scheme with a project number RG94/20.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 2

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 2008. 2

[3] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[5] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European conference on computer vision*, pages 109–122. Springer, 2014. 5

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020. 3

[7] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: generative co-training for generative adversarial networks with limited data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 499–507, 2022. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 3

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2

[11] Jiaxing Huang, Kaiwen Cui, Dayan Guan, Aoran Xiao, Fangneng Zhan, Shijian Lu, Shengcai Liao, and Eric Xing. Masked generative adversarial networks are data-efficient generation learners. *Advances in Neural Information Processing Systems*, 35:2154–2167, 2022. 2

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 5, 7

[14] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 6

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Ali Koksal and Shijian Lu. Rf-gan: A light and reconfigurable network for unpaired image-to-image translation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[17] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2

[18] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 3, 5, 6, 7

[19] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3d model based approach. In *2012 IEEE International Conference on Multimedia and Expo*, pages 333–338. IEEE, 2012. 2

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 1, 5

[22] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 5

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[24] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*, pages 117–132. Springer, 2018. 2

[25] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018. 2

[26] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019. 2, 5, 6

[27] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018. 2

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 4

[29] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 1, 5, 7

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[32] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 5, 6, 7

[33] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 6, 7

[34] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 4

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 6

[38] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 3, 5

[39] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 3

[40] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022. 2

[41] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 2

[42] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Kaiwen Cui, Aoran Xiao, Shijian Lu, and Chunyan Miao. Bi-level feature alignment for versatile image translation and manipulation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 224–241. Springer, 2022. 2

[43] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Kaiwen Cui, Changgong Zhang, and Shijian Lu. Autoregressive image synthesis with integrated quantization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 110–127. Springer, 2022. 3

[44] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10663–10672, 2022. 2

[45] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 2

[46] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-DETR: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731*, 2021. 3

[47] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 4, 5

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3